KONFERENCA
PORTOROŽ, 15. DO 17. MAJ 2017

Microsoft

Azure Data Lake Analytics - ADLA

Tomaž Kaštrun

e: tomaz.kastrun@gmail.com

t:@tomaz_tsql

Thanks to all the sponsors

Microsoft

Reach us with #ntk17

KONFERENCA

About

BI Developer and data analyst

SQL Server, SAS, R, Python, C#, SAP, SPSS

15years experience MSSQL, DEV, BI, DM

Spar ICS Austria, Spar Slovenija

Avid coffee drinker

Bicycle junkie

http://tomaztsql.wordpress.com

tomaz.kastrun@gmail.com

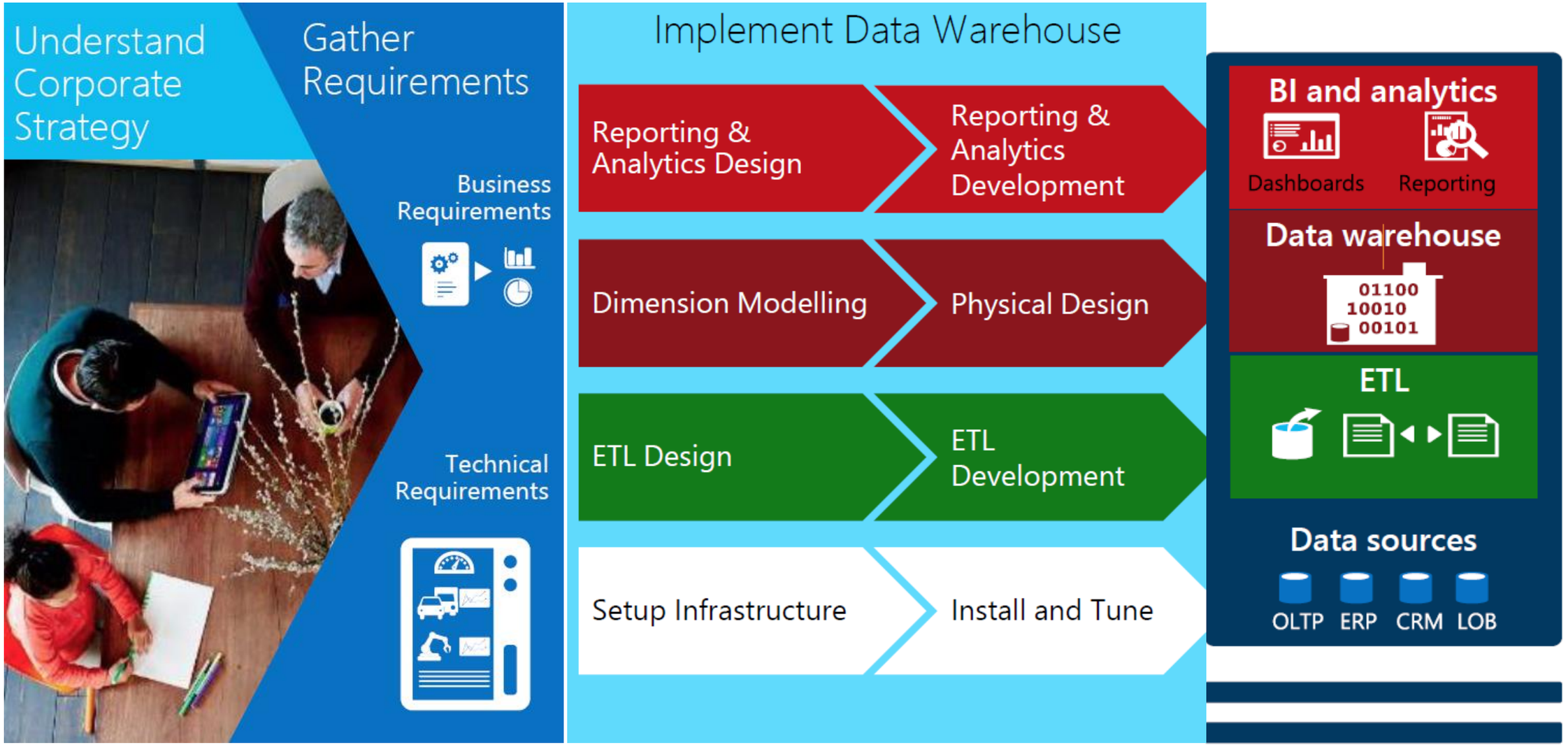
@tomaz_tsql

/in/tomaztsql

http://github.com/tomaztk

https://mvp.microsoft.com/PublicProfile/5002196

Traditional Data Warehouse approach

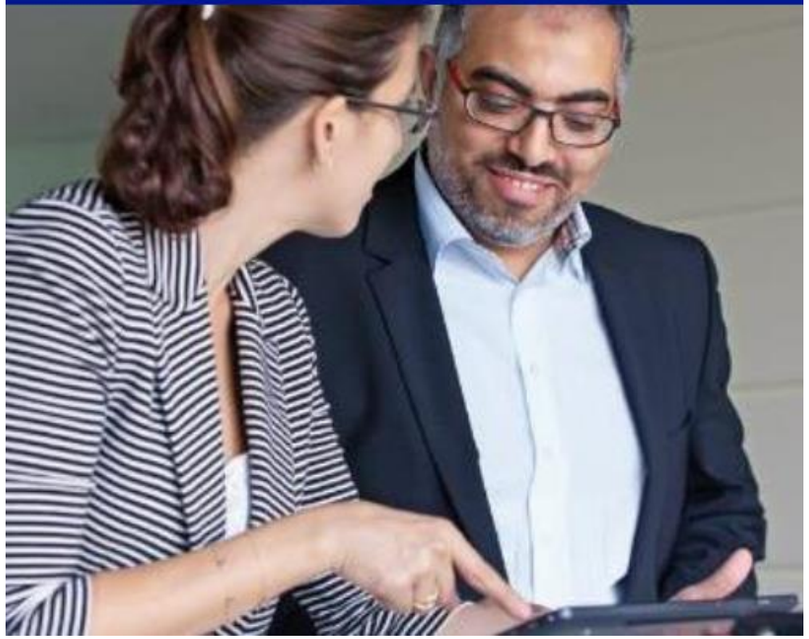


Introducting Azure Data Lake

Big Data made easy

Analytics on any data, any size

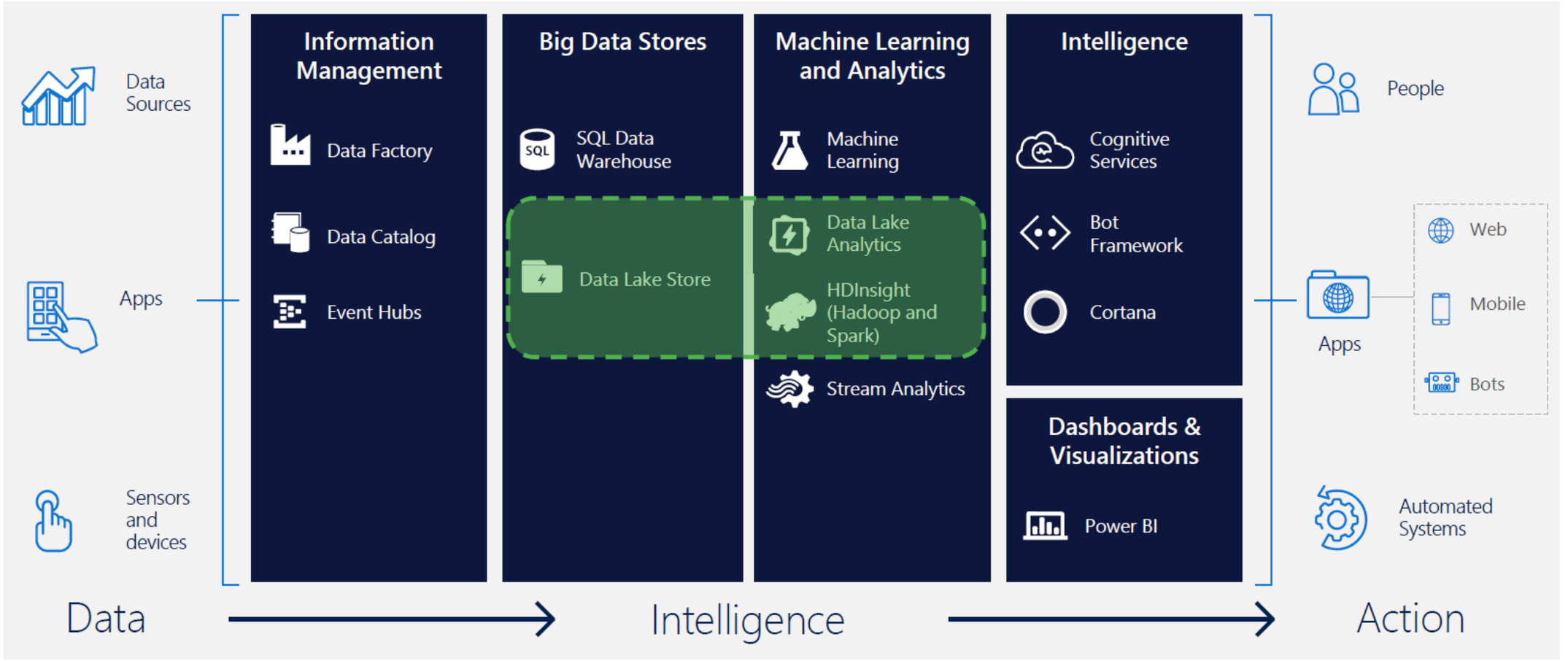
All users productive on day one



Ready for your enterprise

Components of Azure Data Lake

Part of Cortana Intelligence Suite

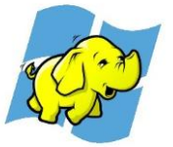


The Data Lake approach

ADL vs./+ HDInsight

1) „Like“ HDFS + YARN + HDInsight
2) Pay as you go (don’t provision)
3) Specify node count (parallelism) at job submission time
4) Optimized storage for big data analytics and workloads
5) Based on Azure Active directory identities
6) No limits on account size, file size or number of files
7) Redundancy: ? ☺
8) Pricing: (next slide)

1) HDFS + YARN
2) Provisioning by clusters
3) Specify clusters of n nodes
4) General object store for variety of storage scenarios
5) Based on shared access signature key authentication
6) Specific limitis
7) Redundancy: LRS, GRS, ZRS, RA-GRS
8) Pricing: (next slide)



cloudera

Further Reading: https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-comparison-with-blob-storage

ADL vs./+ HDInsight

Data Lake Store

REGION: East US 2

PRICING TYPE: Pay-as-you-go

Storage used

5 TB = $199.68/MO

Read Transactions

100000 Transaction units (10,000 transactions) × $0.004 Per 10,000 transactions = $400.00/MO

Write Transactions

100000 Transaction units (10,000 transactions) × $0.05 Per 10,000 transactions = $5,000.00/MO

Delete Transactions are free.

Sub-total $5,599.68/MO

Storage

REGION: West US

TYPE: Block blob

PRICING TIER:

DATA REDUNDANCY:

Storage

REGION: West US

TYPE: Table and Queue

PRICING TIER: Basic

DATA REDUNDANCY: RA-GRS

Capacity

50 TB = $5,140.48/MO

Storage transactions

1000000 Transaction units (10,000 transactions) × $0.00036 Per unit = $360.00/MO

Sub-total $5,500.48/MO

TB

Sub-total $90.75/MO

Why ADLA?
Characteristics of Big Data Analytics

Processing of any type of data

Allows using custom algorithms

Scales efficiently to any size

ADLA

- Enables customers to leverage existing experience with C#, SQL & PowerShell
- Offers convenience, efficiency, automatic scale, and management in a "job service" form factor

Why ADLA?
Characteristics of Big Data Analytics

Start in seconds
Scale instantly
Pay per job

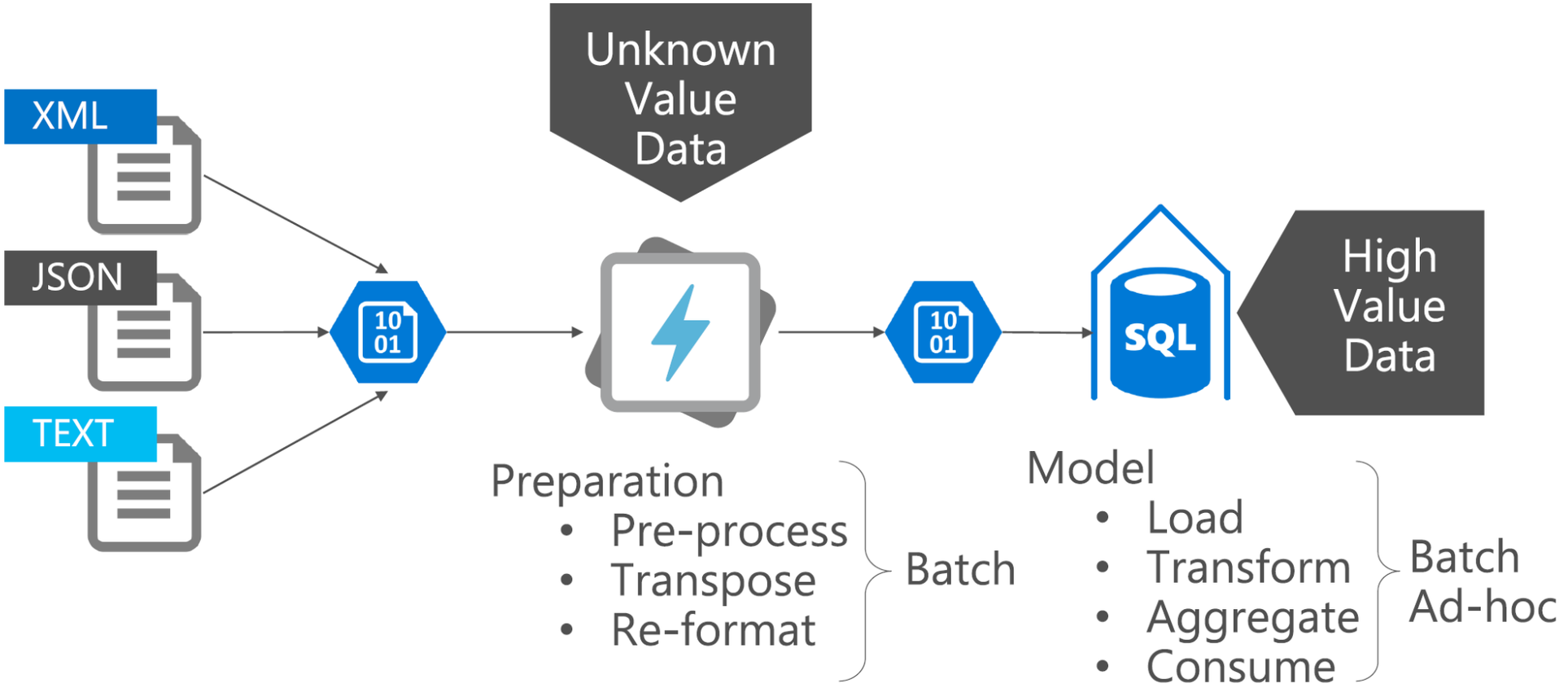
Develop massively parallel programs with simplicity

Debug and optimize your Big Data programs with ease

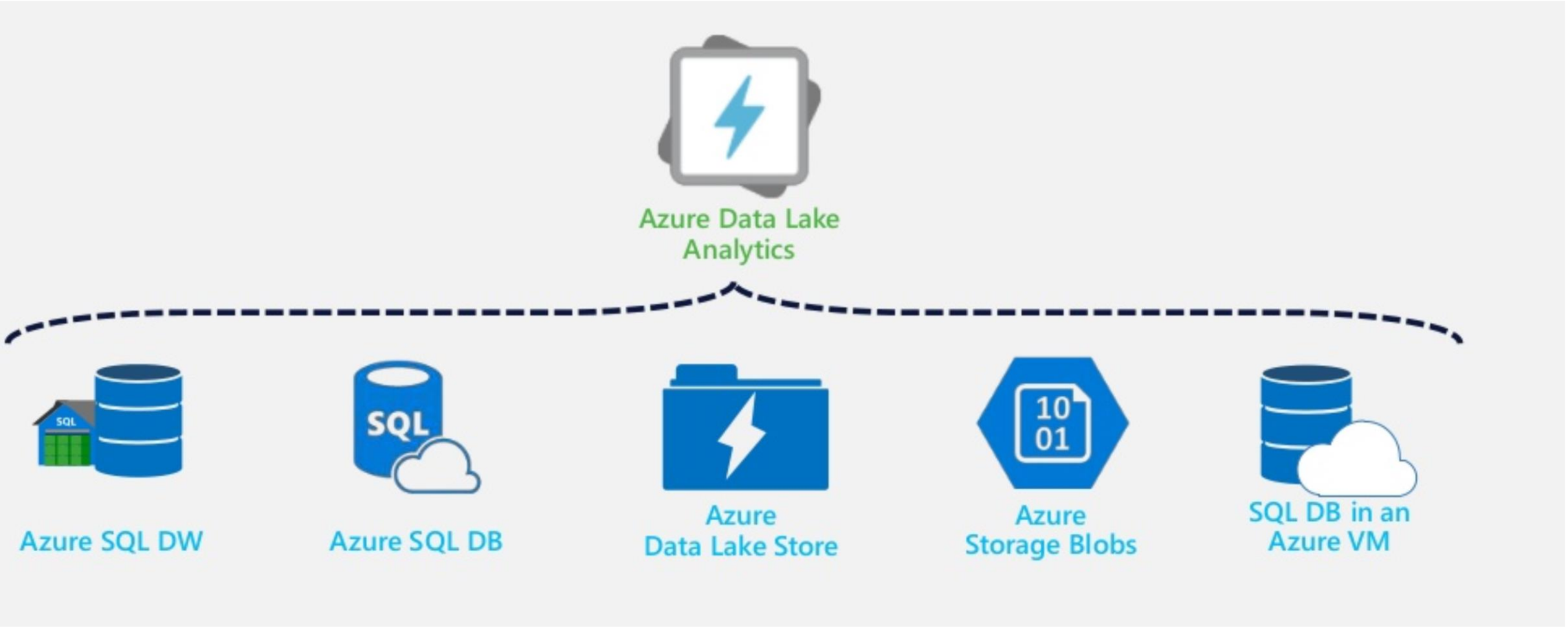
Virtualize your analytics

Enterprise-grade security, auditing and support

ADLS and SQL/DW/PowerBI/…

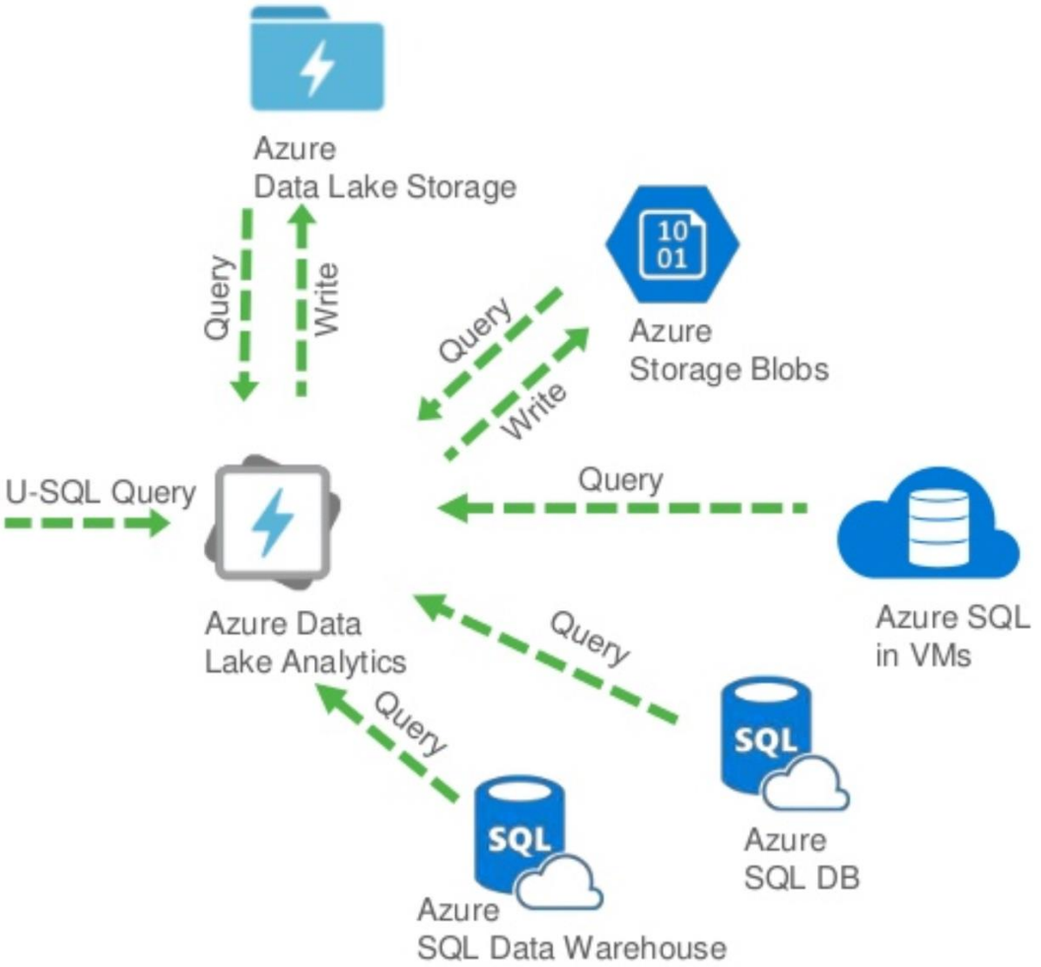


ADLA work across all data



Querying data where it is

No data movement across network (between stores)
No need for single data store
Single view of data irrespective of physical location
Minimizing multiple data copies
Single query language for all data
Each data store maintains its own sovereignty
Pushing SQL expressions to remote SQL sources



U-SQL Origins

SCOPE – Microsoft's internal Big Data Language

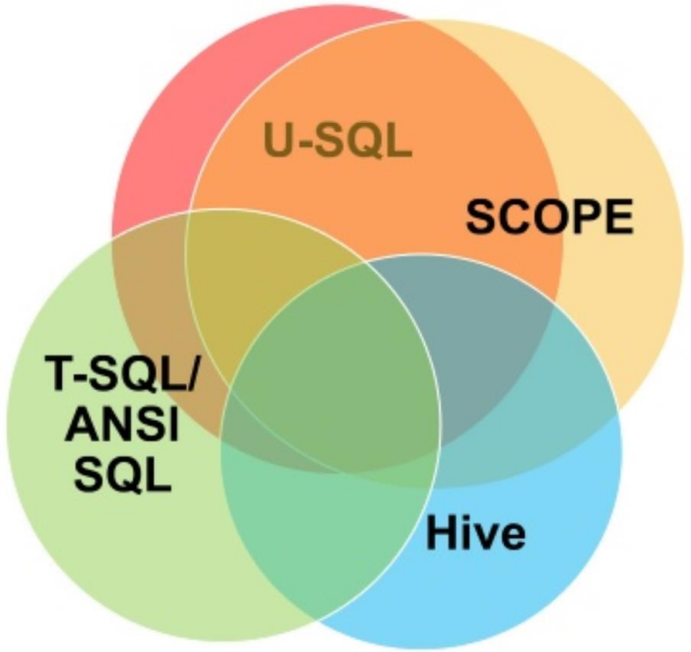
- SQL + C# integration Model
- Optimization and scaling model
- Runs 100.000 jobs daily

HIVE

- Compley data types (Maps, Arrays)
- Data format alignment for text files

ANSI SQL / T-SQL

- Native SQL Language
- Capabilities - procedures, functions, data types,Windowing functions, meta data model,…



U-SQL Language concept

Expression-flow programming style

- Functional lambda expression
- Composable and globally optimizable

Operates on unstructured and structured data

- Schema on read over files
- Relational metadata objects (e.g.table, database)

Extensible from ground up

- Type system is based on C#
- Expression language IS C#
- User-defined functions (U-SQL and C#)
- User-defined aggregators (C#)
- User-defined operators - UDO (C#)

U-SQL provides the parallelization and scale-out framework for usercode

- EXTRACTORS, OUTPUTTER, PROCESSOR, REDUCERS, COMBINER, APPLIERS

Federates query across distributed data sources

```
REFERENCE MyDB.MyAssembly;

CREATE TABLE T( cid int, first_order DateTime
              , last_order DateTime, order_count int
              , order_amount float );

@o = EXTRACT oid int, cid int, odate DateTime, amount float
     FROM "/input/orders.txt"
     USING Extractors.Csv();

@c = EXTRACT cid int, name string, city string
     FROM "/input/customers.txt"
     USING Extractors.Csv();

@j = SELECT c.cid, MIN(o.odate) AS firstorder
           , MAX(o.date) AS lastorder, COUNT(o.oid) AS ordercnt
           , AGG<MyAgg.MySum>(c.amount) AS totalamount
     FROM @c AS c LEFT OUTER JOIN @o AS o ON c.cid == o.cid
     WHERE c.city.StartsWith("New")
        && MyNamespace.MyFunction(o.odate) > 10
     GROUP BY c.cid;

OUTPUT @j TO "/output/result.txt"
USING new MyData.Write();

INSERT INTO T SELECT * FROM @j;
```

Extending U-SQL (with C#/.NET, Python, R)

C# Expressions (in SELECT expressions)

User Defined functions – UDF

User Defined Aggregates – UDAGGS

User Defined Operators – UDO (UDO can be back-stabbing)

Working with assemblies and libraries (C#, NuGet)

Referencing .Net Framework Assemblies

R / Python for extending to predictive analytics

Cognitive assemblies

Powershell scripts

U-SQL Analytics

Windowing expressions

- Over partition by *clause*
- Order by *clause*
- Row *clause*

Windowing aggregate functions

- ANY_VALUE, AVG, COUNT, MAX, MIN, SUM
- STDEV, STDEVP, VAR, VARP

Analytics functions

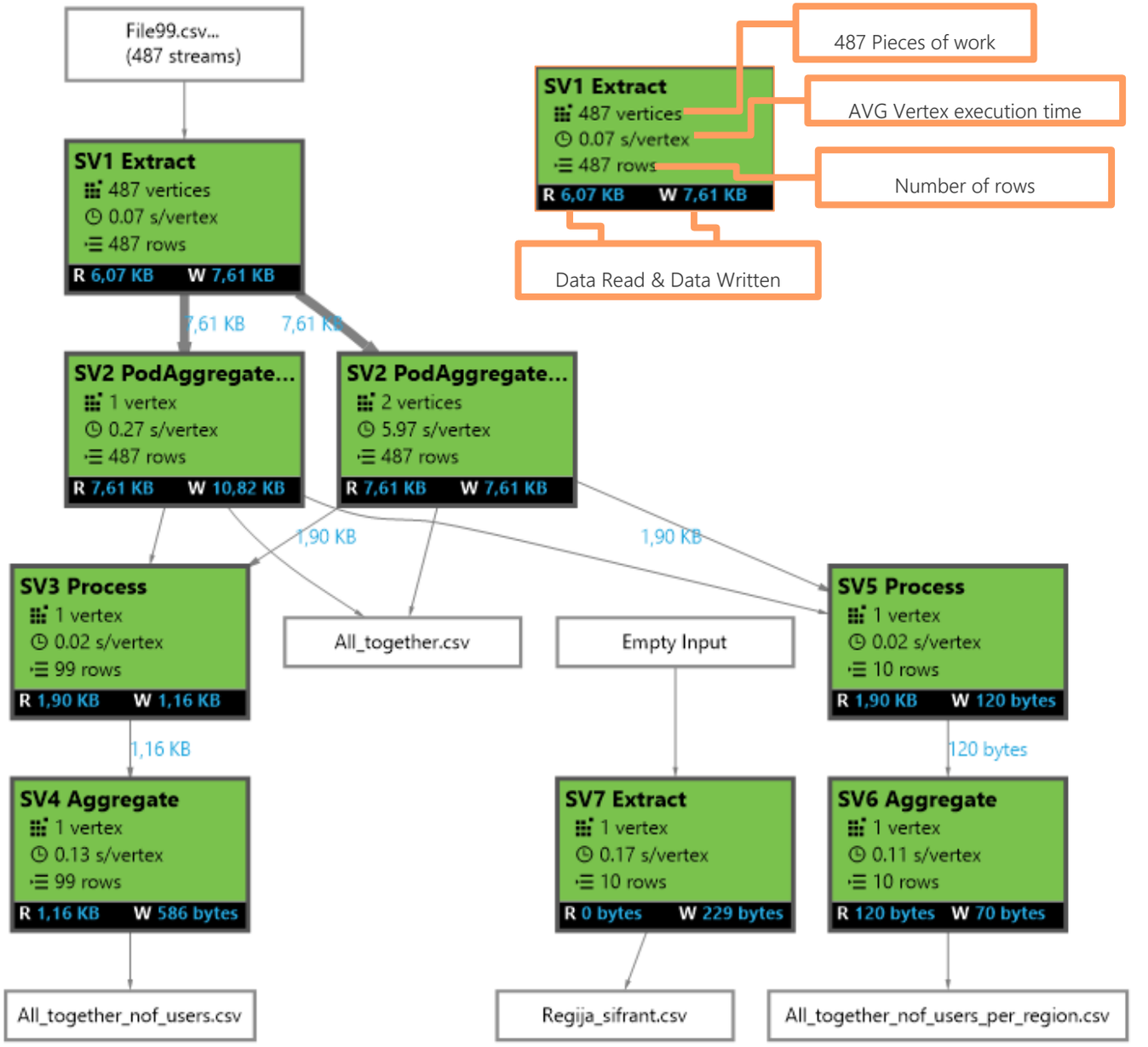
- CUME_DIST, FIST_VALUE, LAST_VALUE, PERCENTILE_CONT, PERCENTILE_DISC, LEAD, LAG, PER_RANK

Ranking functions

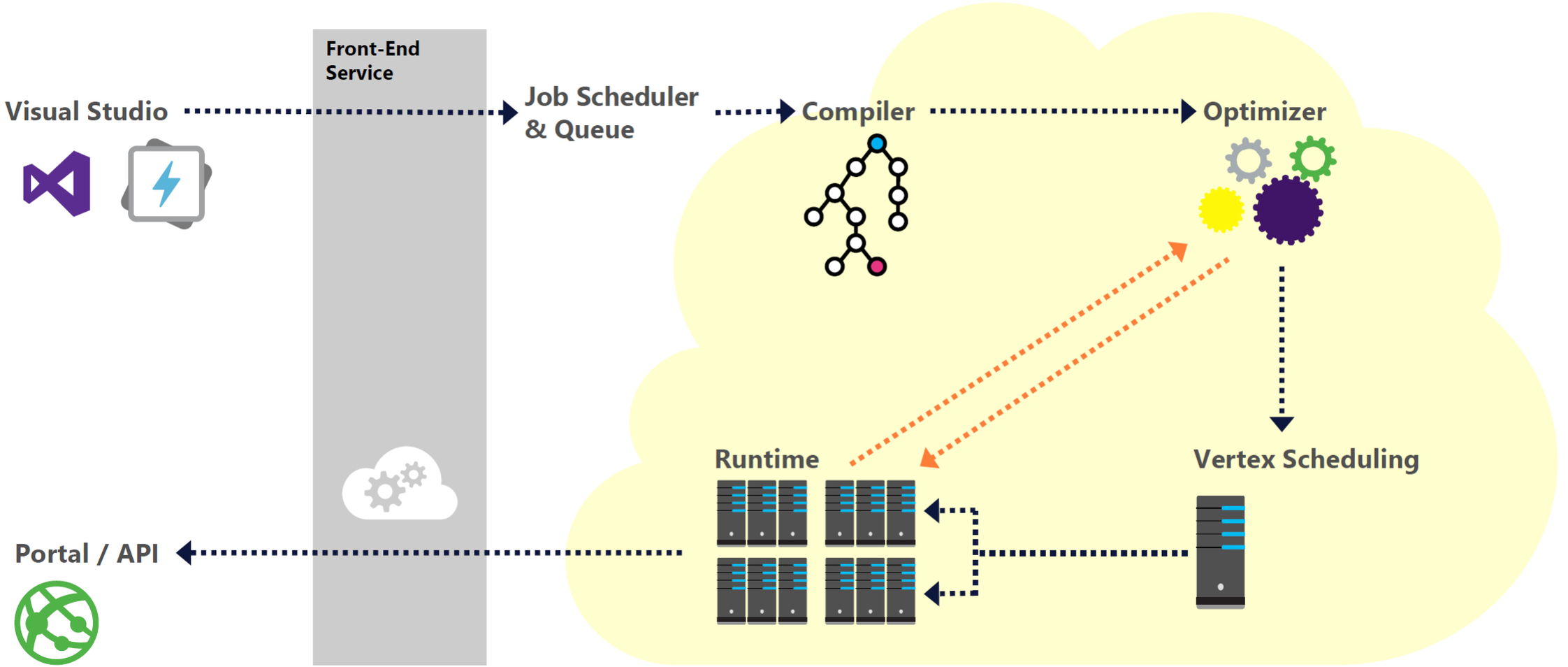
- DENSE_RANK, NTILE, RANK, ROW_NUMBER

U-SQL Execution

1) Script as single execution model with automatic „in-lining“ of U-SQL/C# expressions
2) Automatic query plan optimization
3) Parallelization per-job
4) Detailed information on step execution
5) Debugging with execution plans
6) Colors and vertices to identify performance issues



U-SQL Query processing

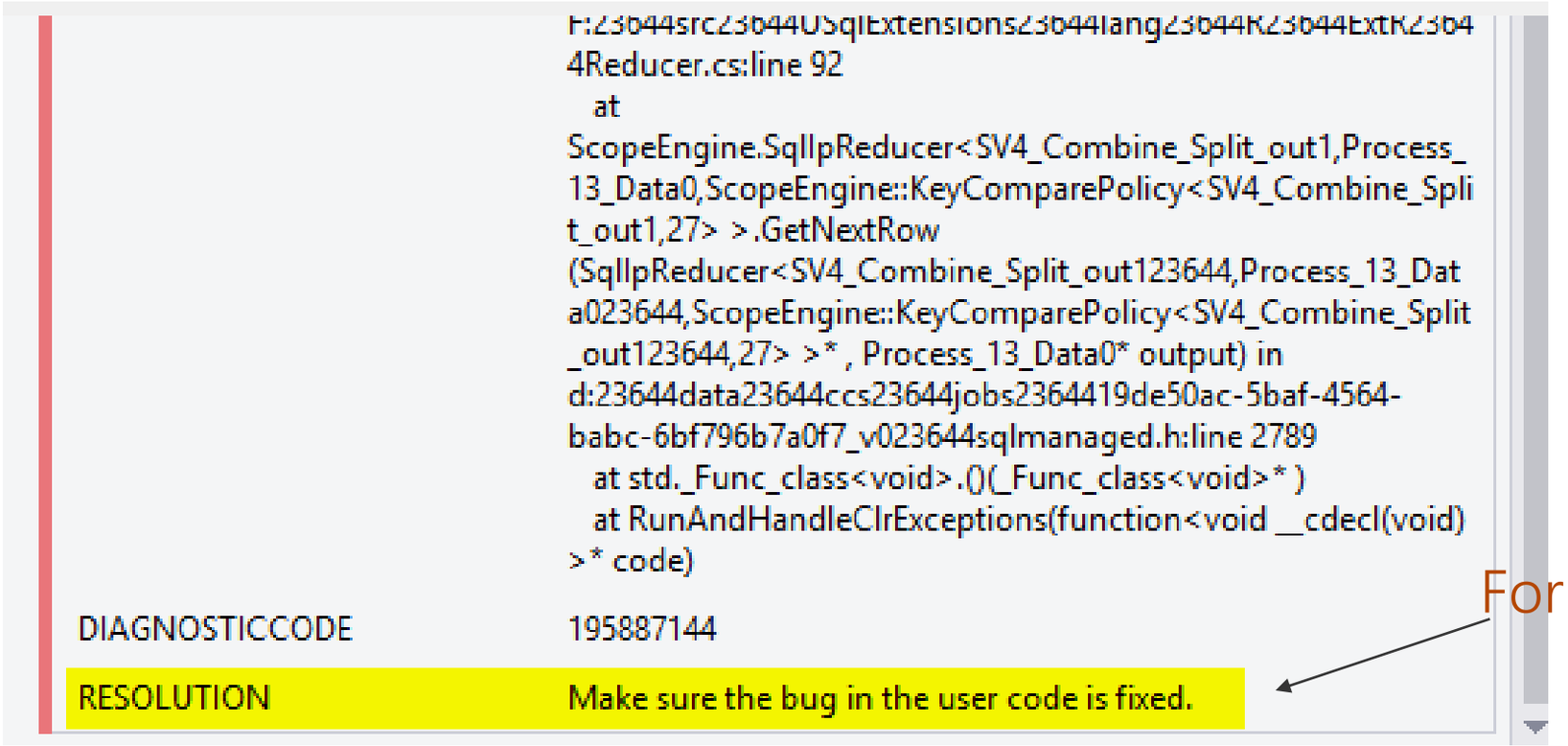


Portal vs. Visual Studio

1) Author U-SQL script
2) Submit / stop U-SQL jobs
3) Visualize and replay progress of a job
4) Overall job management chart and usage stats (compute hours)
5) Provision users who can submit jobs
6) Fine-tune query performance

1) Author U-SQL script (incl. C#)
2) Debug U-SQL and – of course – C# code
3) Submit / stop U-SQL jobs
4) Visualize physical U-SQL query plan
5) Visualize and replay progress of a job
6) Fine-tune query performance
7) Create metadata objects
8) Browse metadata catalog

U-SQL Debugging



For the winners ☺

ADLA billing

Pay for the compute hours for your queries

Pay for storage separately

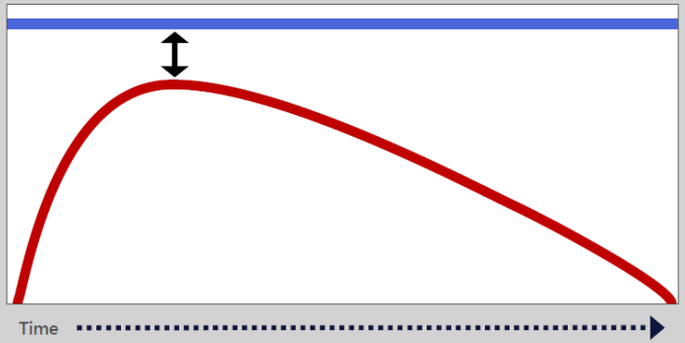
ADLAU allocation

Example:

Allocating 10 ADLAU for 10 minutes job
Cost: 10 min * 10 ADLAU = 100 ADLAU minutes
(formula: query_hours * parallelism * price/hour)

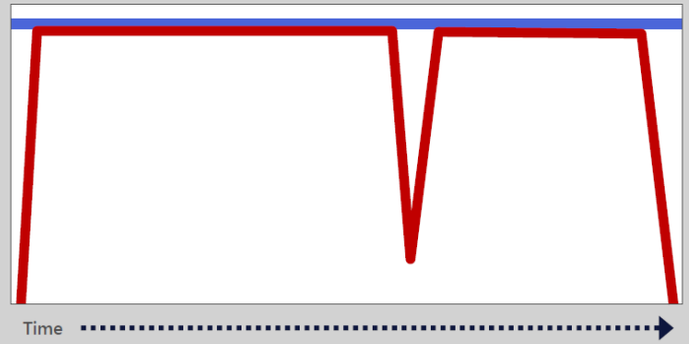
USAGE	GA PRICE (STARTING JANUARY 1ST, 2017)*
ADLAU	$2 / hour
Completed Job	Free

Over-allocation



- Consider using fewer ADLAUs
- You are paying for the area under the blue line
- You are only using the area under the red line

Under-allocation



- Consider using more ADLAUs

Blue line: allocated
Red line: running

Additional read: https://blogs.msdn.microsoft.com/azuredatalake/2016/10/12/understanding-adl-analytics-unit/

Finally … demo(s)

Microsoft

KONFERENCA

ADLA & U-SQL Recap

Processing Big data

Unifies natively T-SQL's declaratively and C# extensibility

Unbound to schema and relational model

Enterprise ready, Security and auditing support

Python and R support

Increase productivity and be „agile"

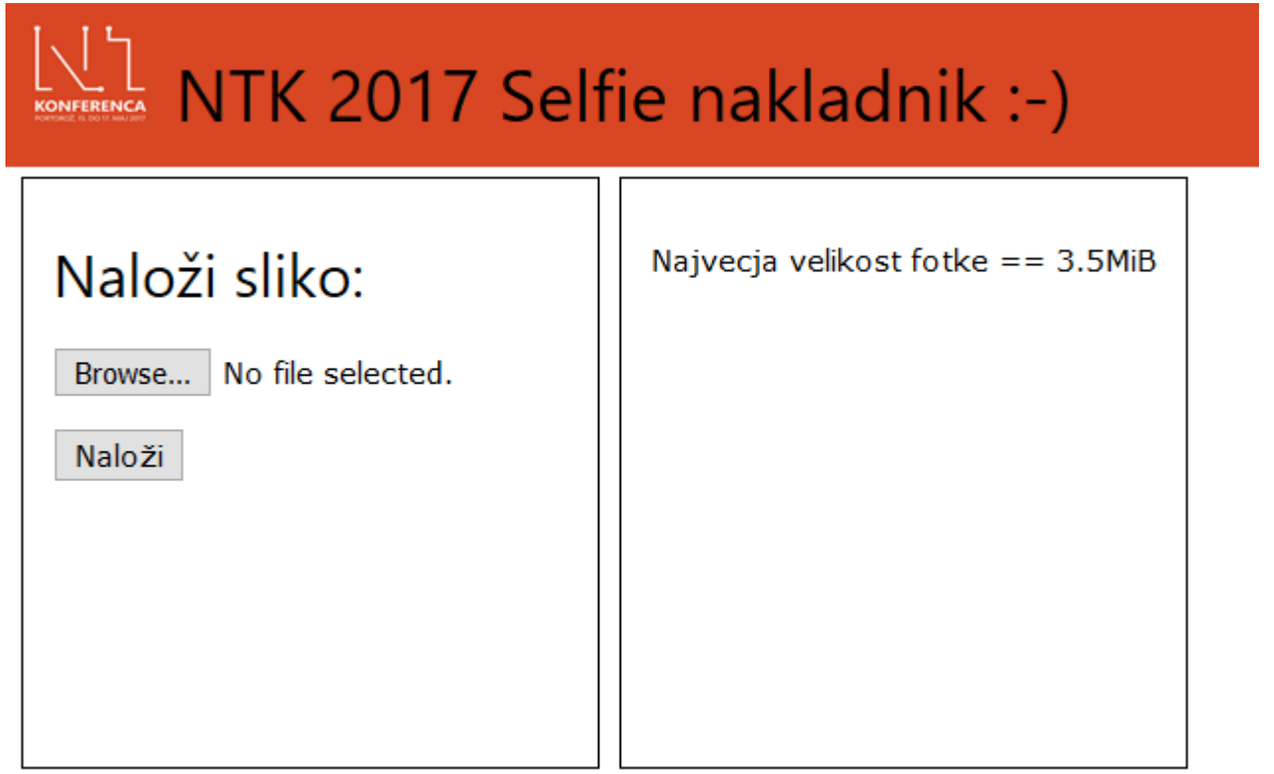
Selfie Time!

Microsoft

KONFERENCA

What to do?

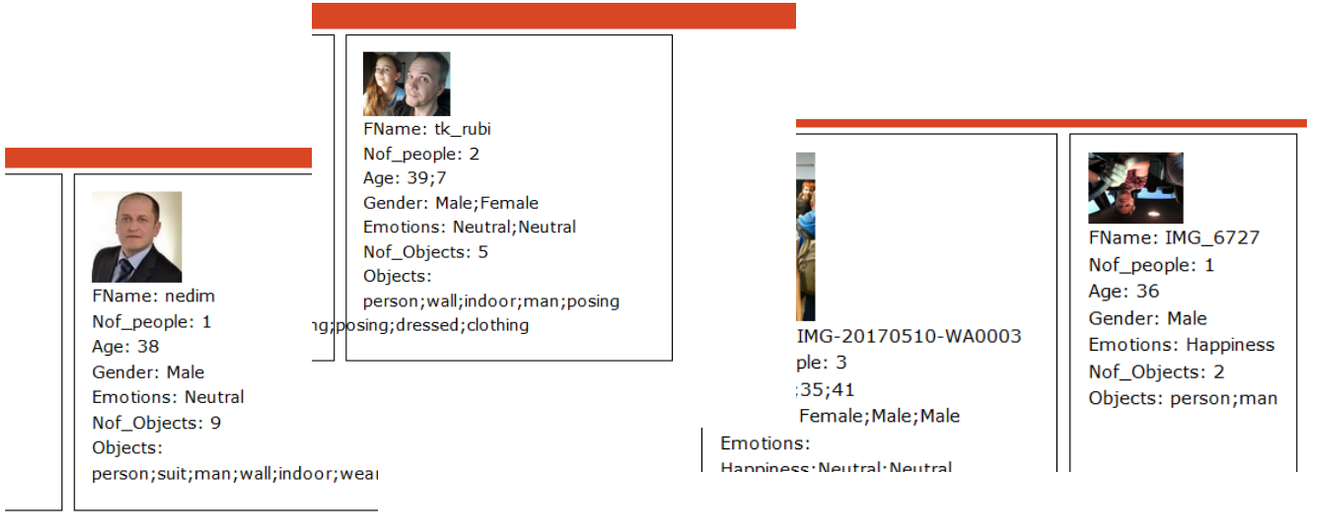
1) Make a Selfie on your mobile phone! ☺ (max 3.5Mib)
2) Upload pic http://ntk2017fotke.azurewebsites.net/



What to do?

Now … executing some Powershell code!

Check: http://ntk2017fotke.azurewebsites.net/flafla.php



Use cases from the field

Microsoft

KONFERENCA

Business use-cases

Worked on:

1) Energy consumption
2) Hotels booking based on image recommendation

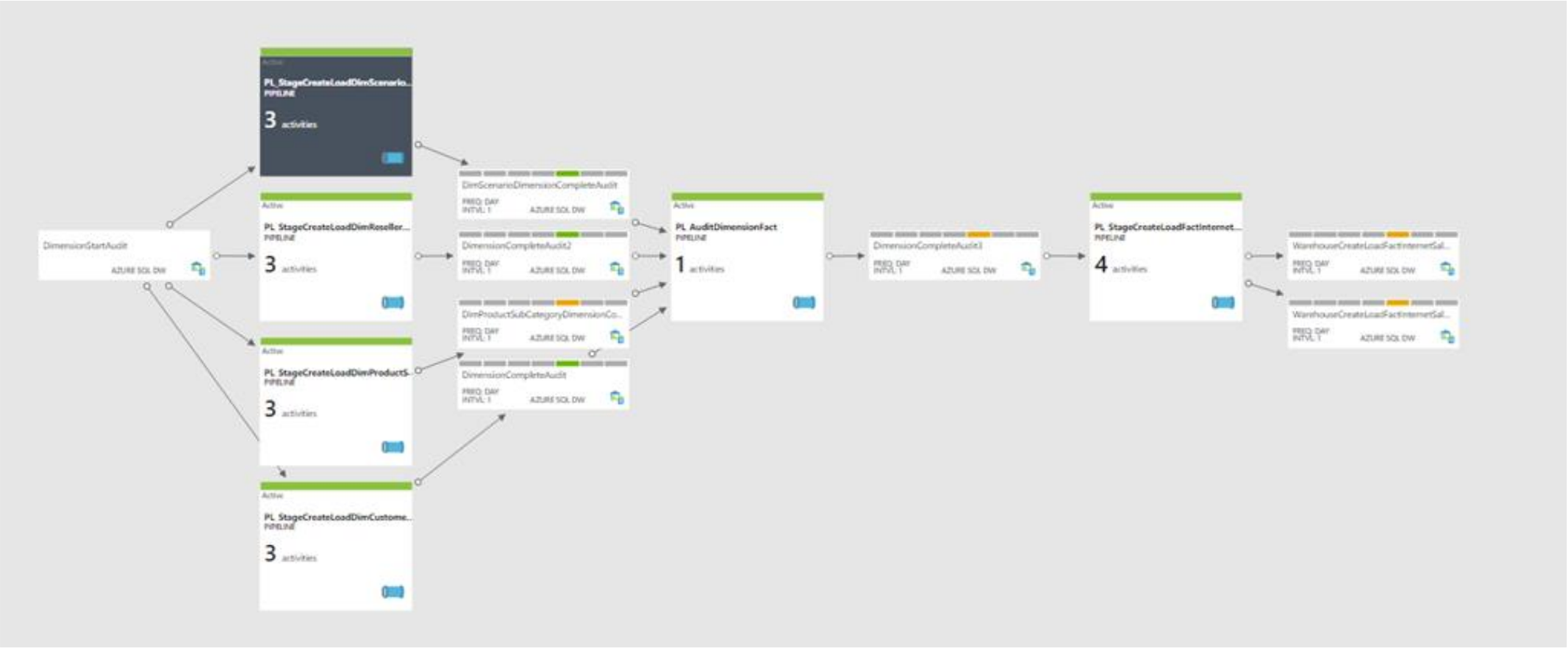
By Microsoft:

1) Digital crime unit – analyze complex attack patterns
2) Image processing – Large-scale image feature extraction and classification using custom code
3) Shopping recommendation – Complex pattern analysis and prediction over shopping records using proprietary algorithms

Hotels / Travel agencies / Acommodations

Energy consumption / IoT

Typical data lake + data factory pipeline



Additional resources

Blogs and community page:

- http://usql.io(U-SQL Github)
- http://blogs.msdn.microsoft.com/mrys/
- http://blogs.msdn.microsoft.com/azuredatalake/
- https://channel9.msdn.com/Search?term=U-SQL#ch9Search

Documentation and articles:

- http://aka.ms/usql_reference
- https://azure.microsoft.com/en-us/documentation/services/data-lake-analytics/
- https://msdn.microsoft.com/en-us/magazine/mt614251

ADL forums and feedback

- http://aka.ms/adlfeedback
- https://social.msdn.microsoft.com/Forums/azure/en-US/home?forum=AzureDataLake
- http://stackoverflow.com/questions/tagged/u-sql

Thank you!
The beer is on me!

Microsoft

KONFERENCA

Community Event!

Reserve the date! 9th December 2017

SQL Saturday Slovenia is
coming to Ljubljana!

SQLSATURDAY

Microsoft

Reach us with #ntk17

KONFERENCA

Microsoft

KONFERENCA

PORTOROŽ, 15. DO 17. MAJ 2017